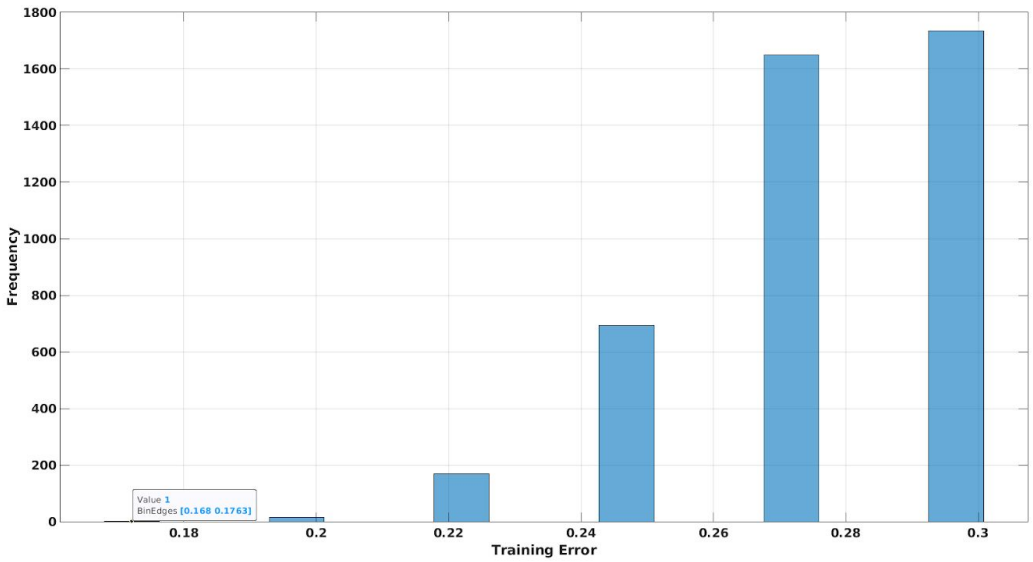
Cross Validation in Feature Selection

For finding the minimum Test error, i had tried multiple times. For this I tried a threshold criteria to stop if TestError < 0.4. These results are arranged as top five sorted Test Error.

Feature Set	Kfold Train Error	Test Error
[5,6,7,8,9,11,12,14]	0.45	0.424
[3,5,6,7,8,9,11,12,14]	0.475	0.424
[1,2,4,5,6,7,8,9,10,11,12,14,15]	0.45	0.426
[3,4,5,6,7,8,9,11,12,14]	0.475	0.426
[2,3,4,5,7,8,9,10,11,12,14]	0.4	0.427

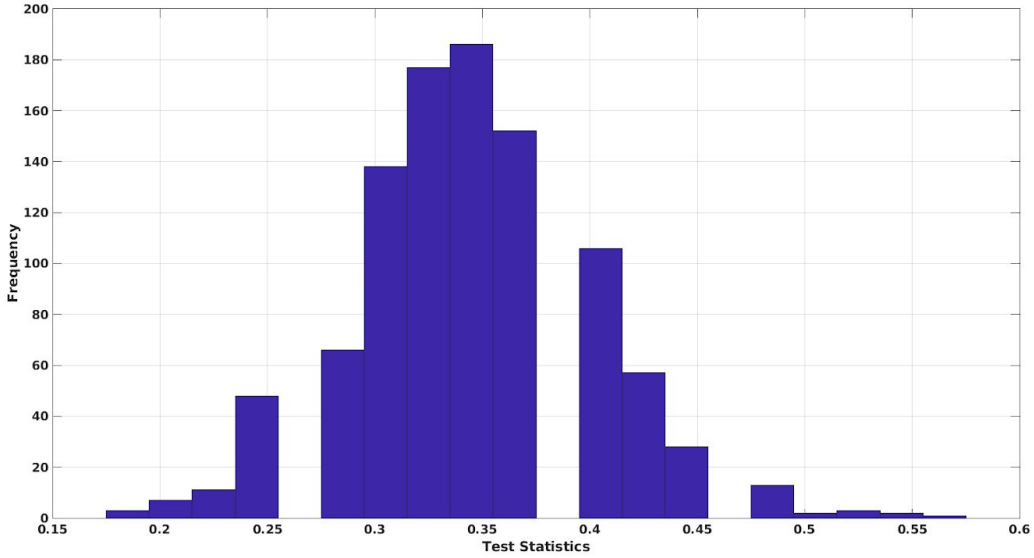


Even though the lowest training error which i found was about 0.17, the test error on those features were greater than the test error reported in table above.

In my view there exist some discrepancy, which is caused by overfitting of the partitioning function. Even though with 32767 features subset/permutation, the possibility for a random subset that fits the data is very high. We should conduct a permutation test to check if Y is independent on features or not.

Permutation test

As the null hypothesis states (X and Y being independent), in that case cross validation error should be around 0.5. The reason for this is that results are not because of overfitting, random permutation of the predicted label should give us about 50% error. Whereas, our best cross validation error is about 0.3 and p-value of 0.139 using 1000 permutations.



Any arbitrary interval [0,x] can be selected for region rejection such that

$$p\,(T < x) \leq \alpha = 0.05$$

In our case the null hypothesis is not rejected as p-value is 0.139 which is greater than 0.05 (alpha). Thus we can not find any information regarding X and Y being independent or not.